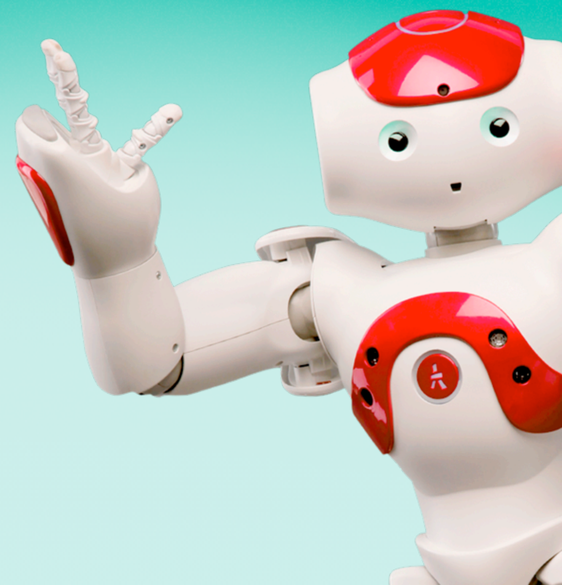
A Web-Based Wizard-of-Oz Platform for Collaborative and Reproducible Human-Robot Interaction Research

Sean O'Connor '26
Department of Computer Science, Bucknell University

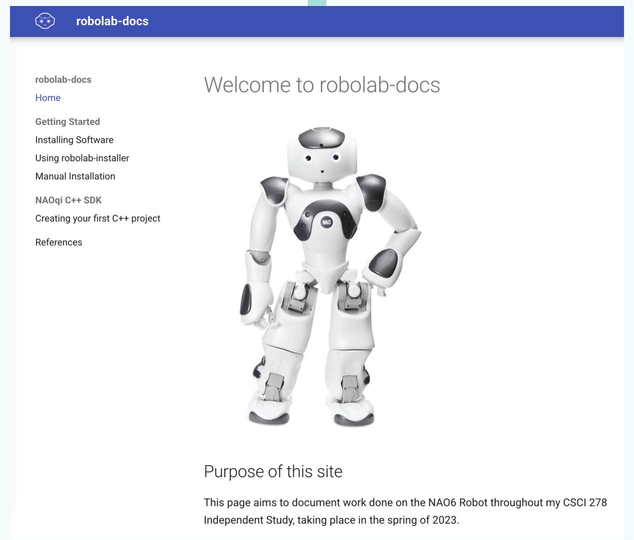
April 20th, 2026

Bucknell UNIVERSITY

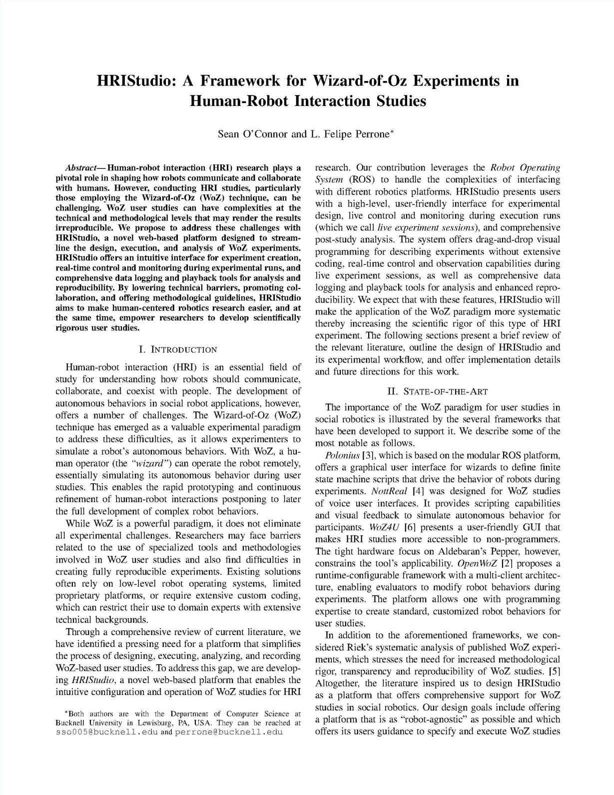


A Web-Based Wizard-of-Oz Platform for Collaborative and Reproducible Human-Robot Interaction Research

2023

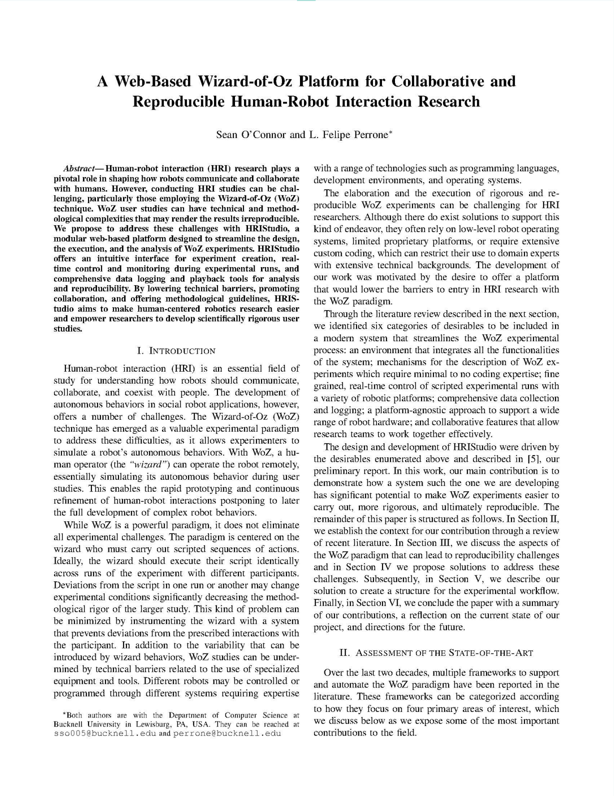


2024


HRIStudio: A Framework for Wizard-of-Oz Experiments in Human-Robot Interaction Studies

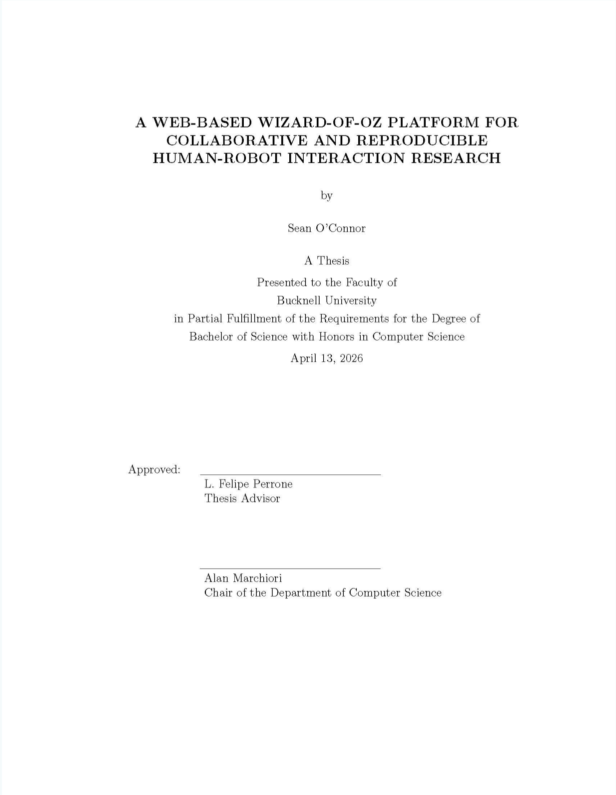
Sean O'Connor and L. Felipe Perrone*

2025


A Web-Based Wizard-of-Oz Platform for Collaborative and Reproducible Human-Robot Interaction Research

Sean O'Connor and L. Felipe Perrone*

2026


A WEB-BASED WIZARD-OF-OZ PLATFORM FOR COLLABORATIVE AND REPRODUCIBLE HUMAN-ROBOT INTERACTION RESEARCH

by

Sean O'Connor

A Thesis

Presented to the Faculty of

Bucknell University

in Partial Fulfillment of the Requirements for the Degree of

Bachelor of Science with Honors in Computer Science

April 13, 2026

Approved:

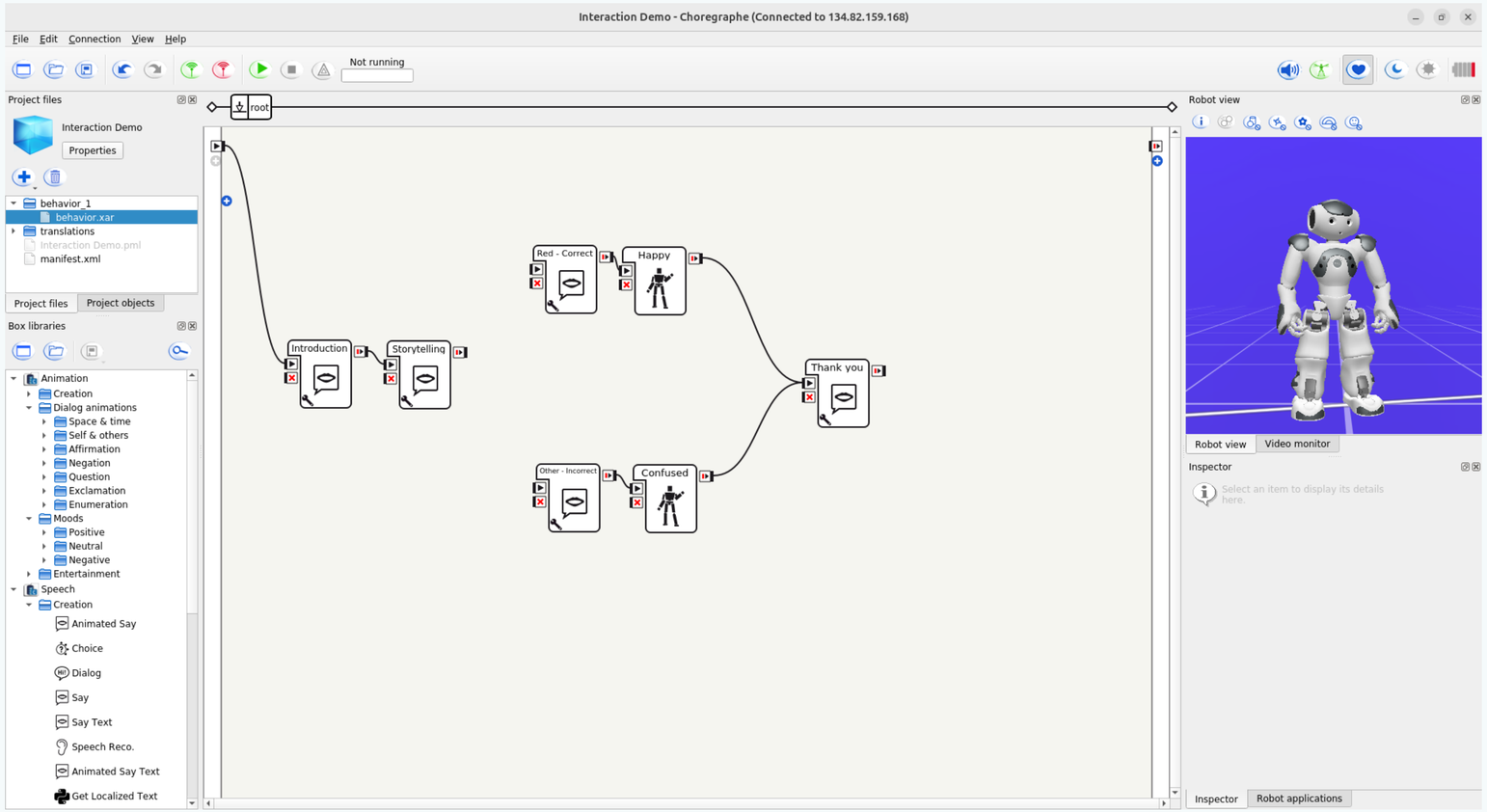
L. Felipe Perrone
Thesis Advisor

Alan Marchiori
Chair of the Department of Computer Science

Meet the NAO

- ~2 foot tall humanoid robot. Designed for research and education.
- One of the most widely deployed social robots worldwide.
- Released 2006. Company changed hands 4–5 times since.

How do we use the NAO?



Choregraphe by Aldebaran

Analyzing Choregraphe

Social Robotics

We study how people interact with physical robots in everyday environments.

- Social robots are entering schools, hospitals, homes
- How people respond shapes how we build them
- *NAO: the most widely used humanoid in social robotics research*

Controlled experiments let us get real answers- which raises the question of how to run them.

Wizard-of-Oz Studies

We need a framework to run these experiments, so we use the **Wizard-of-Oz** framework.

A hidden researcher - the *wizard* - controls the robot in real time, while the *participant* believes it is autonomous.

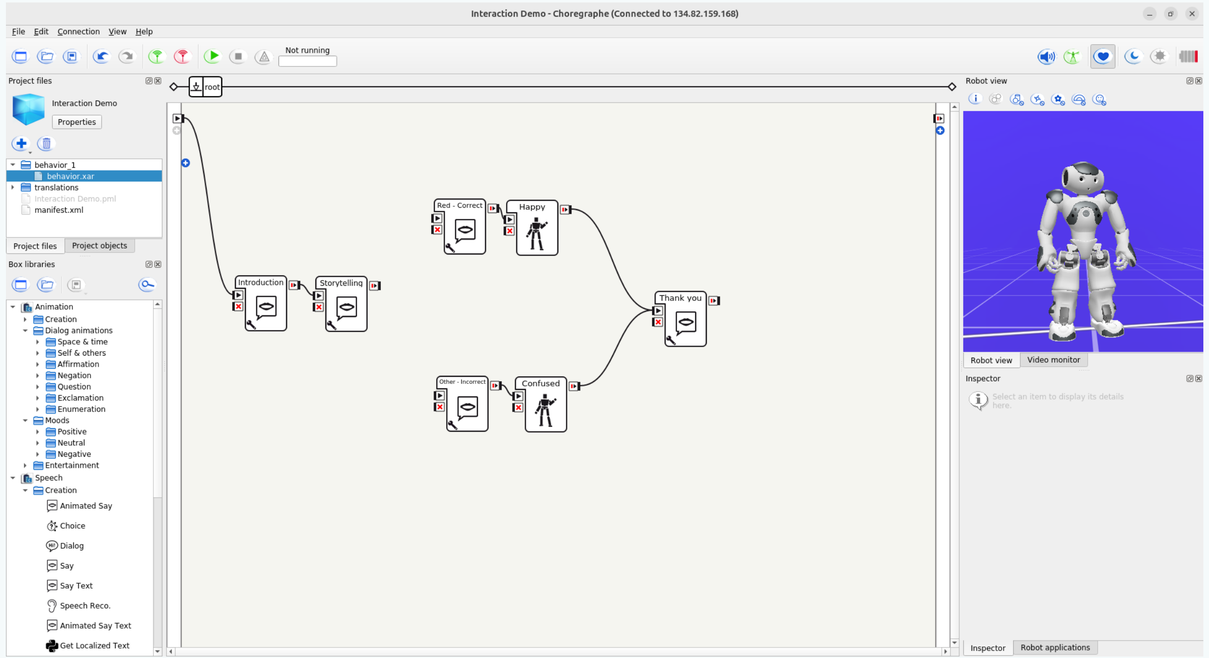
Why?

- Prototype behavior without full programmed autonomy
- Study authentic human response

Choregraphe: A General Purpose Tool

Choregraphe can control the robot well, but it's not designed for the kinds of studies run in social robotics research.

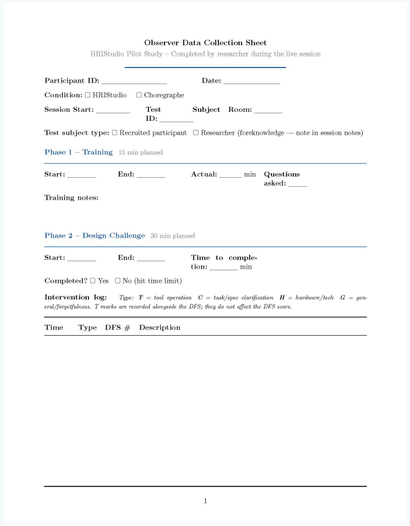
- Flow-based IDE, not a study tool
- Programming experience required
- No study structure, no protocol
- Trigger a behavior mid-trial?
  - Manually hunt through a flow diagram



How do we analyze?

Choregraphe handles control- but what about everything else?

- No logging. Nothing is captured when the trial ends.
- Data must be manually recorded in real time.
- Studies are comprised of much more than just code- where do all the documents go?



Observer Data Collection Sheet

HRIStudio Pilot Study – Completed by researcher during the live session

Participant ID: _____ **Date:** _____

Condition: ☐ HRIStudio ☐ Choregraphe

Session Start: _____ **Test ID:** _____ **Subject** **Room:** _____

Test subject type: ☐ Recruited participant ☐ Researcher (foreknowledge — note in session notes)

Phase 1 – Training 15 min planned

Start: _____ **End:** _____ **Actual:** _____ min **Questions asked:** _____

Training notes:

Phase 2 – Design Challenge 30 min planned

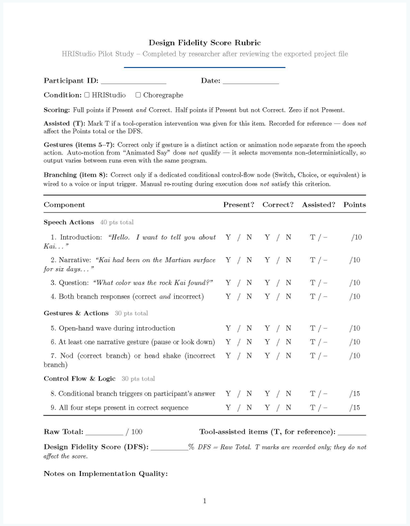
Start: _____ **End:** _____ **Time to completion:** _____ min

Completed? ☐ Yes ☐ No (hit time limit)

Intervention log: *Type: T = tool operation C = task/spec clarification H = hardware/tech G = general/forgetfulness. T marks are recorded alongside the DFS; they do not affect the DFS score.*

Time	Type	DFS #	Description

1



Design Fidelity Score Rubric

HRIStudio Pilot Study – Completed by researcher after reviewing the exported project file

Participant ID: _____ **Date:** _____

Condition: ☐ HRIStudio ☐ Choregraphe

Scoring: Full points if Present *and* Correct. Half points if Present but not Correct. Zero if not Present.

Assisted (T): Mark T if a tool-operation intervention was given for this item. Recorded for reference — does *not* affect the Points total or the DFS.

Gestures (items 5–7): Correct only if gesture is a distinct action or animation node separate from the speech action. Auto-motion from "Animated Say" does *not* qualify — it selects movements non-deterministically, so output varies between runs even with the same program.

Branching (item 8): Correct only if a dedicated conditional control-flow node (Switch, Choice, or equivalent) is wired to a voice or input trigger. Manual re-routing during execution does *not* satisfy this criterion.

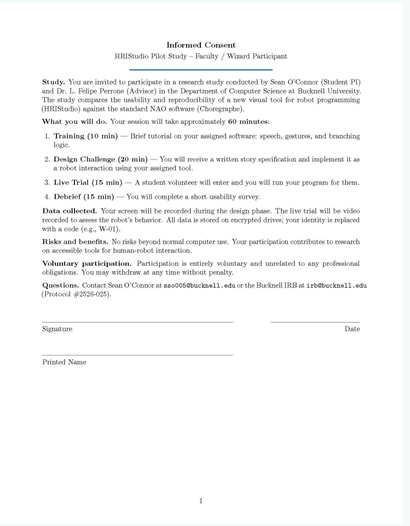
Component	Present?	Correct?	Assisted?	Points
Speech Actions 40 pts total				
1. Introduction: *"Hello. I want to tell you about Kai. . ."*	Y / N	Y / N	T / –	/10
2. Narrative: *"Kai had been on the Martian surface for six days. . ."*	Y / N	Y / N	T / –	/10
3. Question: *"What color was the rock Kai found?"*	Y / N	Y / N	T / –	/10
4. Both branch responses (correct *and* incorrect)	Y / N	Y / N	T / –	/10
Gestures & Actions 30 pts total				
5. Open-hand wave during introduction	Y / N	Y / N	T / –	/10
6. At least one narrative gesture (pause or look down)	Y / N	Y / N	T / –	/10
7. Nod (correct branch) or head shake (incorrect branch)	Y / N	Y / N	T / –	/10
Control Flow & Logic 30 pts total				
8. Conditional branch triggers on participant's answer	Y / N	Y / N	T / –	/15
9. All four steps present in correct sequence	Y / N	Y / N	T / –	/15

Raw Total: _____ / 100 **Tool-assisted items (T, for reference):** _____

Design Fidelity Score (DFS): _____% *DFS = Raw Total. T marks are recorded only; they do not affect the score.*

Notes on Implementation Quality:

1



Informed Consent

HRIStudio Pilot Study – Faculty / Wizard Participant

Study. You are invited to participate in a research study conducted by Sean O'Connor (Student PI) and Dr. L. Felipe Perrone (Advisor) in the Department of Computer Science at Bucknell University. The study compares the usability and reproducibility of a new visual tool for robot programming (HRIStudio) against the standard NAO software (Choregraphe).

What you will do. Your session will take approximately **60 minutes**:

1. **Training (10 min)** — Brief tutorial on your assigned software: speech, gestures, and branching logic.
2. **Design Challenge (20 min)** — You will receive a written story specification and implement it as a robot interaction using your assigned tool.
3. **Live Trial (15 min)** — A student volunteer will enter and you will run your program for them.
4. **Debrief (15 min)** — You will complete a short usability survey.

Data collected. Your screen will be recorded during the design phase. The live trial will be video recorded to assess the robot's behavior. All data is stored on encrypted drives; your identity is replaced with a code (e.g., W-01).

Risks and benefits. No risks beyond normal computer use. Your participation contributes to research on accessible tools for human-robot interaction.

Voluntary participation. Participation is entirely voluntary and unrelated to any professional obligations. You may withdraw at any time without penalty.

Questions. Contact Sean O'Connor at `sso005@bucknell.edu` or the Bucknell IRB at `irb@bucknell.edu` (Protocol #2526-025).

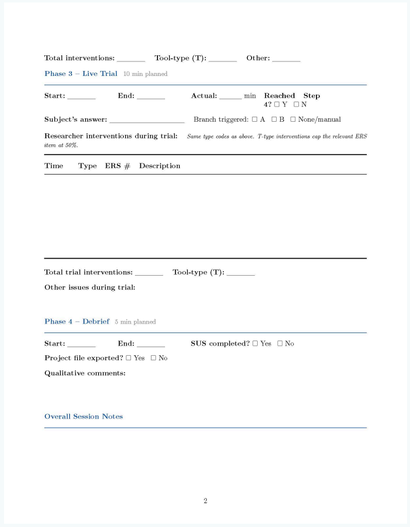
_____ _____

Signature Date

_____

Printed Name

1



Total interventions: _____ **Tool-type (T):** _____ **Other:** _____

Phase 3 – Live Trial 10 min planned

Start: _____ **End:** _____ **Actual:** _____ min **Reached Step 4?** ☐ Y ☐ N

Subject's answer: _____ Branch triggered: ☐ A ☐ B ☐ None/manual

Researcher interventions during trial: *Same type codes as above. T-type interventions cap the relevant ERS item at 50%.*

Time	Type	ERS #	Description

Total trial interventions: _____ **Tool-type (T):** _____

Other issues during trial:

Phase 4 – Debrief 5 min planned

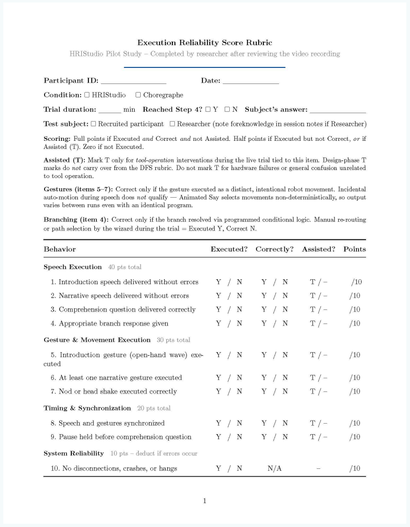
Start: _____ **End:** _____ **SUS completed?** ☐ Yes ☐ No

Project file exported? ☐ Yes ☐ No

Qualitative comments:

Overall Session Notes

2



Execution Reliability Score Rubric

HRIStudio Pilot Study – Completed by researcher after reviewing the video recording

Participant ID: _____ **Date:** _____

Condition: ☐ HRIStudio ☐ Choregraphe

Trial duration: _____ min **Reached Step 4?** ☐ Y ☐ N **Subject's answer:** _____

Test subject: ☐ Recruited participant ☐ Researcher (note foreknowledge in session notes if Researcher)

Scoring: Full points if Executed *and* Correct *and* not Assisted. Half points if Executed but not Correct, *or* if Assisted (T). Zero if not Executed.

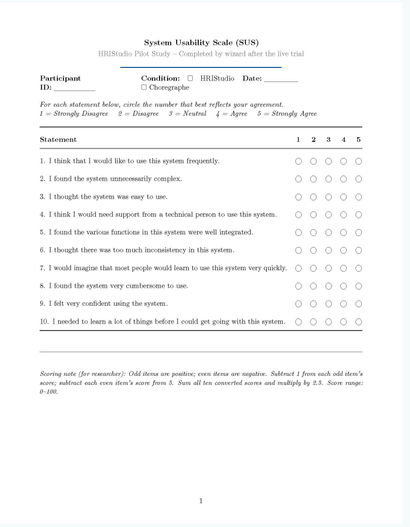
Assisted (T): Mark T only for *tool-operation* interventions during the live trial tied to this item. Design-phase T marks do not carry over from the DFS rubric. Do not mark T for hardware failures or general confusion unrelated to tool operation.

Gestures (items 5–7): Correct only if the gesture executed as a distinct, intentional robot movement. Incidental auto-motion during speech does *not* qualify — Animated Say selects movements non-deterministically, so output varies between runs even with an identical program.

Branching (item 4): Correct only if the branch resolved via programmed conditional logic. Manual re-routing or path selection by the wizard during the trial = Executed Y, Correct N.

Behavior	Executed?	Correctly?	Assisted?	Points
Speech Execution 40 pts total				
1. Introduction speech delivered without errors	Y / N	Y / N	T / –	/10
2. Narrative speech delivered without errors	Y / N	Y / N	T / –	/10
3. Comprehension question delivered correctly	Y / N	Y / N	T / –	/10
4. Appropriate branch response given	Y / N	Y / N	T / –	/10
Gesture & Movement Execution 30 pts total				
5. Introduction gesture (open-hand wave) executed	Y / N	Y / N	T / –	/10
6. At least one narrative gesture executed	Y / N	Y / N	T / –	/10
7. Nod or head shake executed correctly	Y / N	Y / N	T / –	/10
Timing & Synchronization 20 pts total				
8. Speech and gestures synchronized	Y / N	Y / N	T / –	/10
9. Pause held before comprehension question	Y / N	Y / N	T / –	/10
System Reliability 10 pts – deduct if errors occur				
10. No disconnections, crashes, or hangs	Y / N	N/A	–	/10

1



System Usability Scale (SUS)

HRIStudio Pilot Study – Completed by wizard after the live trial

Participant ID: _____ **Condition:** ☐ HRIStudio ☐ Choregraphe **Date:** _____

For each statement below, circle the number that best reflects your agreement.
1 = Strongly Disagree 2 = Disagree 3 = Neutral 4 = Agree 5 = Strongly Agree

Statement	1	2	3	4	5
1. I think that I would like to use this system frequently.	○	○	○	○	○
2. I found the system unnecessarily complex.	○	○	○	○	○
3. I thought the system was easy to use.	○	○	○	○	○
4. I think I would need support from a technical person to use this system.	○	○	○	○	○
5. I found the various functions in this system were well integrated.	○	○	○	○	○
6. I thought there was too much inconsistency in this system.	○	○	○	○	○
7. I would imagine that most people would learn to use this system very quickly.	○	○	○	○	○
8. I found the system very cumbersome to use.	○	○	○	○	○
9. I felt very confident using the system.	○	○	○	○	○
10. I needed to learn a lot of things before I could get going with this system.	○	○	○	○	○

Scoring note (for researcher): Odd items are positive; even items are negative. Subtract 1 from each odd item's score; subtract each even item's score from 5. Sum all ten converted scores and multiply by 2.5. Score range: 0–100.

1

A Known Crisis in HRI Methodology

In 2012, Laurel D. Riek published a paper titled “Wizard of Oz Studies in HRI: A Systematic Review and New Reporting Guidelines.”

- Systematic review of 54 WoZ studies in HRI (2001 - 2011).
- Only 5.4% of studies reported pre-experiment wizard training.
- Only 3.7% of studies reported measuring wizard error.

The field lacks tools to **constrain wizard behavior** and **enforce standardized protocols**, making studies nearly impossible to replicate.

What about the rest of the field?

Other Solutions Came Before

	Low technical barrier	High technical barrier
More rigorous	?	Polonius, OpenWoZ Flexible, but requires programming
Less rigorous	WoZ4U Accessible, but platform-specific No methodological rigor	Choregraphe Requires programming experience or training No methodological rigor

Distilling the problems

Two Problems

Problem 1:

The Accessibility Problem

Existing tools require substantial programming expertise, which prevents domain experts from conducting independent social robotics studies.

Problem 2:

The Reproducibility Problem

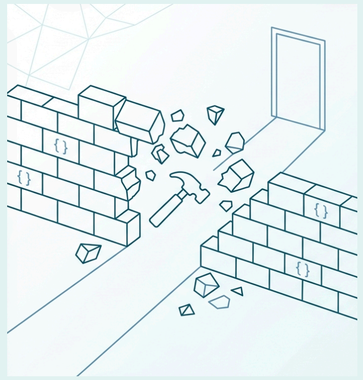
Without structured logging and protocol enforcement, **experiment execution varies across participants and wizards** in ways that are difficult to detect or control after the fact.

Two Research Questions

Research Question 1:

Accessibility

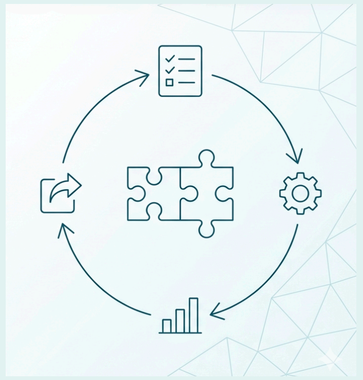
Can researchers **without programming expertise** use a purpose-built platform to design and conduct Wizard-of-Oz studies?



Research Question 2:

Reproducibility

Can a WoZ-specific platform enable more **faithful and reliable execution** of study design compared to existing tooling?



How can we fix this?

Three Design Principles

Design Principle 1:

Hierarchical Specification

Study designs are **structured data**, not code.

Study → Experiment → Step → Action.

Shareable, analyzable, version-controllable.

Design Principle 2:

Event-driven Execution

Wizard clicks through.

Behavior fires completely, and in order.

No searching for the correct next step.

Design Principle 3:

Modular Architecture

Swap the NAO plugin for another robot platform.

Study design persists.

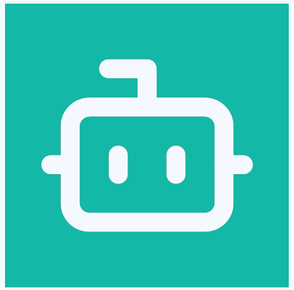
Robot-agnostic by design.

Existing Solutions

	Low technical barrier	High technical barrier
More rigorous	?	Polonius OpenWoZ Flexible, but requires programming
Less rigorous	WoZ4U Accessible, but platform-specific No methodological rigor	Choregraphe Requires programming experience or training No methodological rigor

Existing Solutions

	Low technical barrier	High technical barrier
More rigorous	The answer? **HRIStudio**	Polonius OpenWoZ Flexible, but requires programming
Less rigorous	WoZ4U Accessible, but platform-specific No methodological rigor	Choregraphe Requires programming experience or training No methodological rigor



HRIStudio

The Experiment Designer

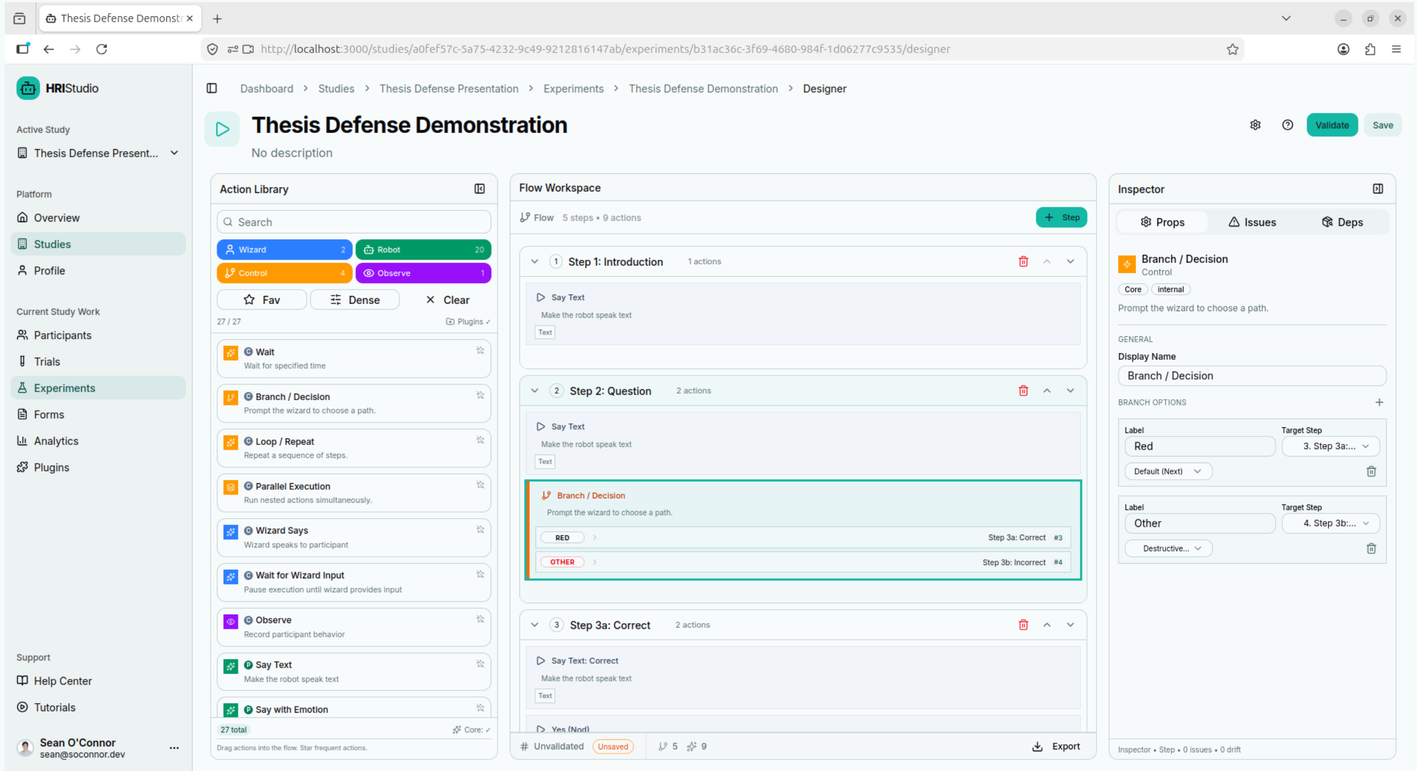
HRIStudio's *Experiment Designer* allows a non-programmer to drag-and-drop predefined actions into a script.

The hierarchical specification is enforced here:

A study has experiments.

↳An experiment has steps.

↳A step has actions.

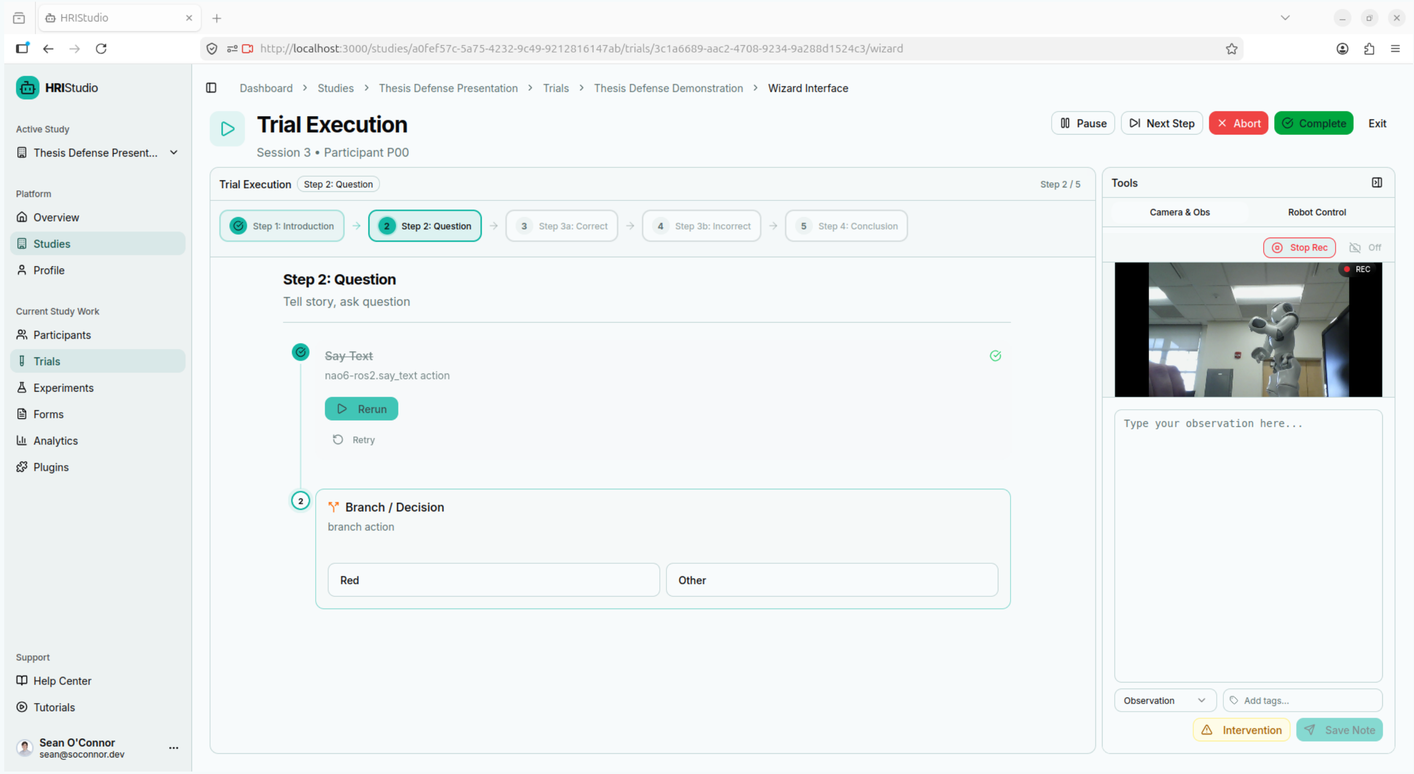


The Execution Interface

The wizard clicks through an experiment's script-following the designed order, but allowing for human intervention if needed.

Video and audio are recorded for future viewing, time-synced to trial events.

Any deviations from the experiment script are logged and marked for later analysis.

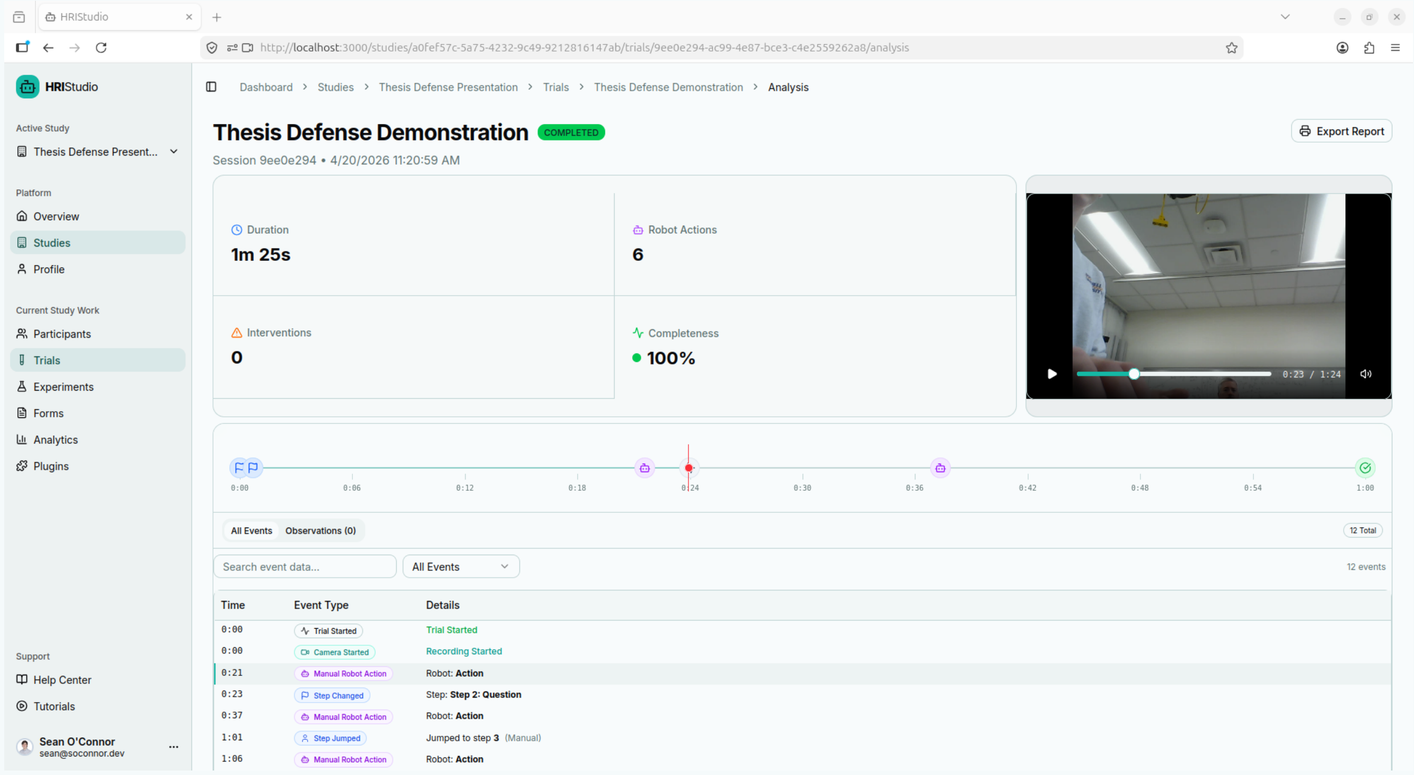


The Trial Report

The *Trial Report* shows a time-synced view of a trial's events.

Each event is timestamped and marked, allowing the researcher to analyze what happened, when it happened- so they can more easily figure out why.

Any deviation from the experiment script is easily viewable here.



The Study

Designing a Pilot Study

N = 6
Faculty participants

Between-subjects
Isolated conditions

60 minutes
Session length

Tool Training	Design	Trial	Debrief
15 minutes	30 minutes	10 minutes	5 min

The Task

Implement and execute “The Interactive Storyteller” - a standardized social robotics scenario in which the robot narrates a story to a participant, then asks a recall question.

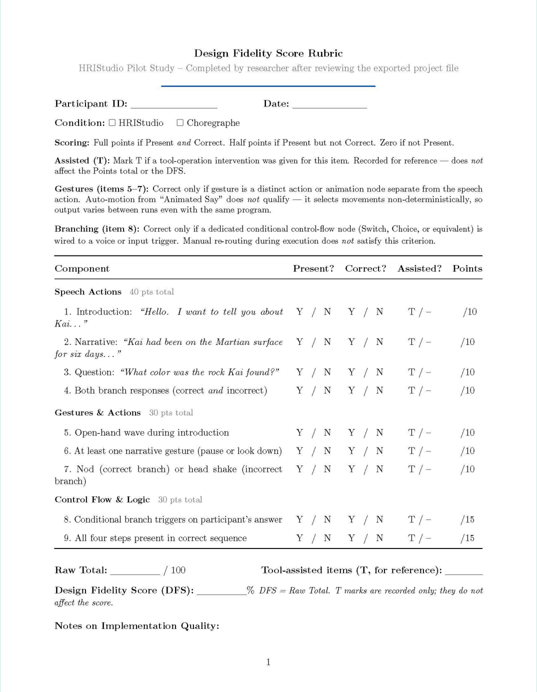
Measuring Success

Measurement Scale 1:

Design Fidelity Score (DFS)

Original measure

What fraction of the intended study design did the participant faithfully implement? Measures how well the platform lets researchers realize their design intent.

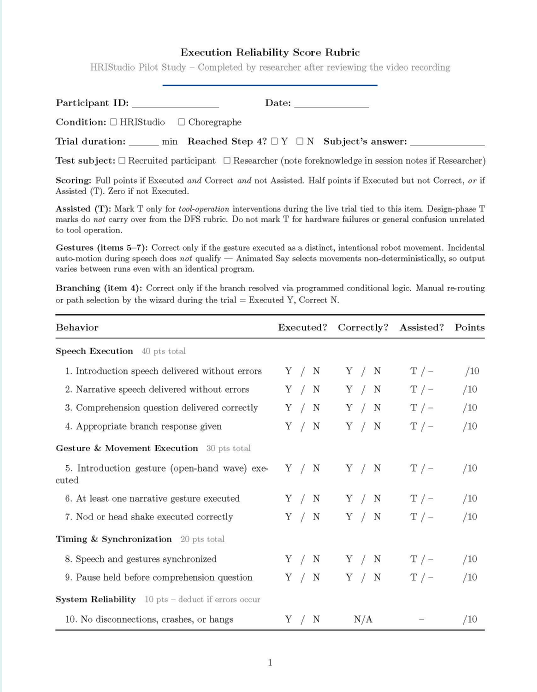

Design Fidelity Score Rubric

Measurement Scale 2:

Execution Reliability Score (ERS)

Original measure

What fraction of planned behaviors executed correctly and on-cue during the live trial? Measures real-time reliability under actual study conditions.

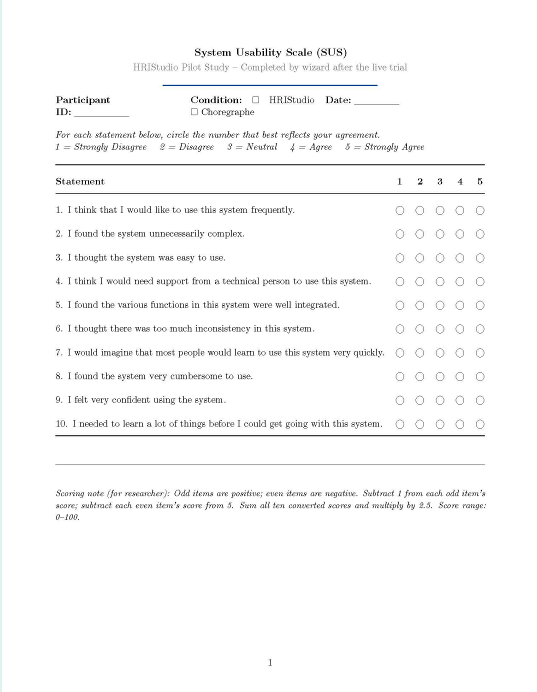

Execution Reliability Score Rubric

Measurement Scale 3:

System Usability Scale (SUS)

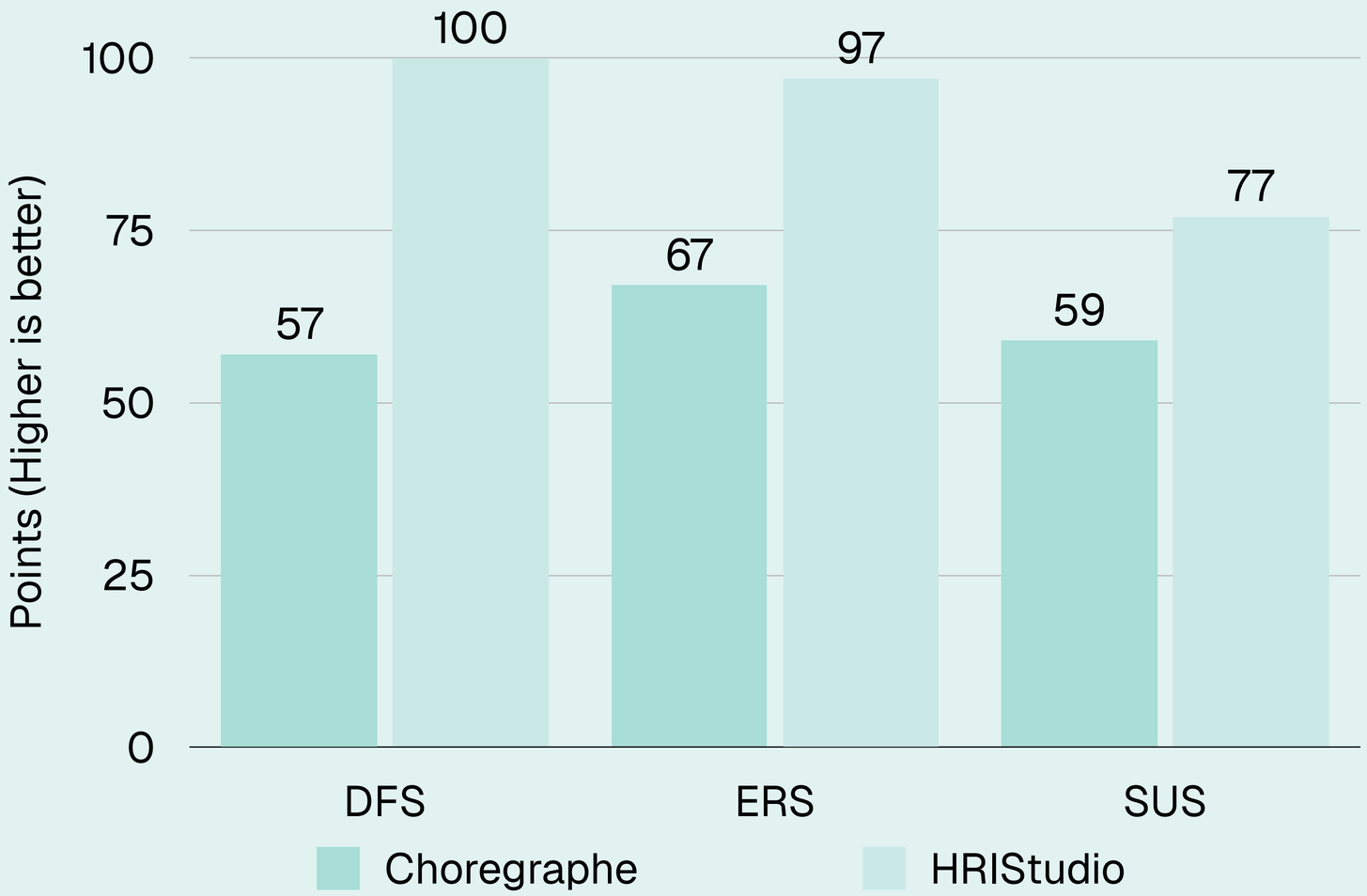
Validated · Benchmark 68

10-item standardized questionnaire measuring perceived ease of use. Scores above 68 indicate above-average usability across industry benchmarks.


System Usability Scale (SUS)

Results

Scoring the Trials



Design Fidelity Scale

+43 points

Perfect design, every time.

Execution Reliability Score

+30 points

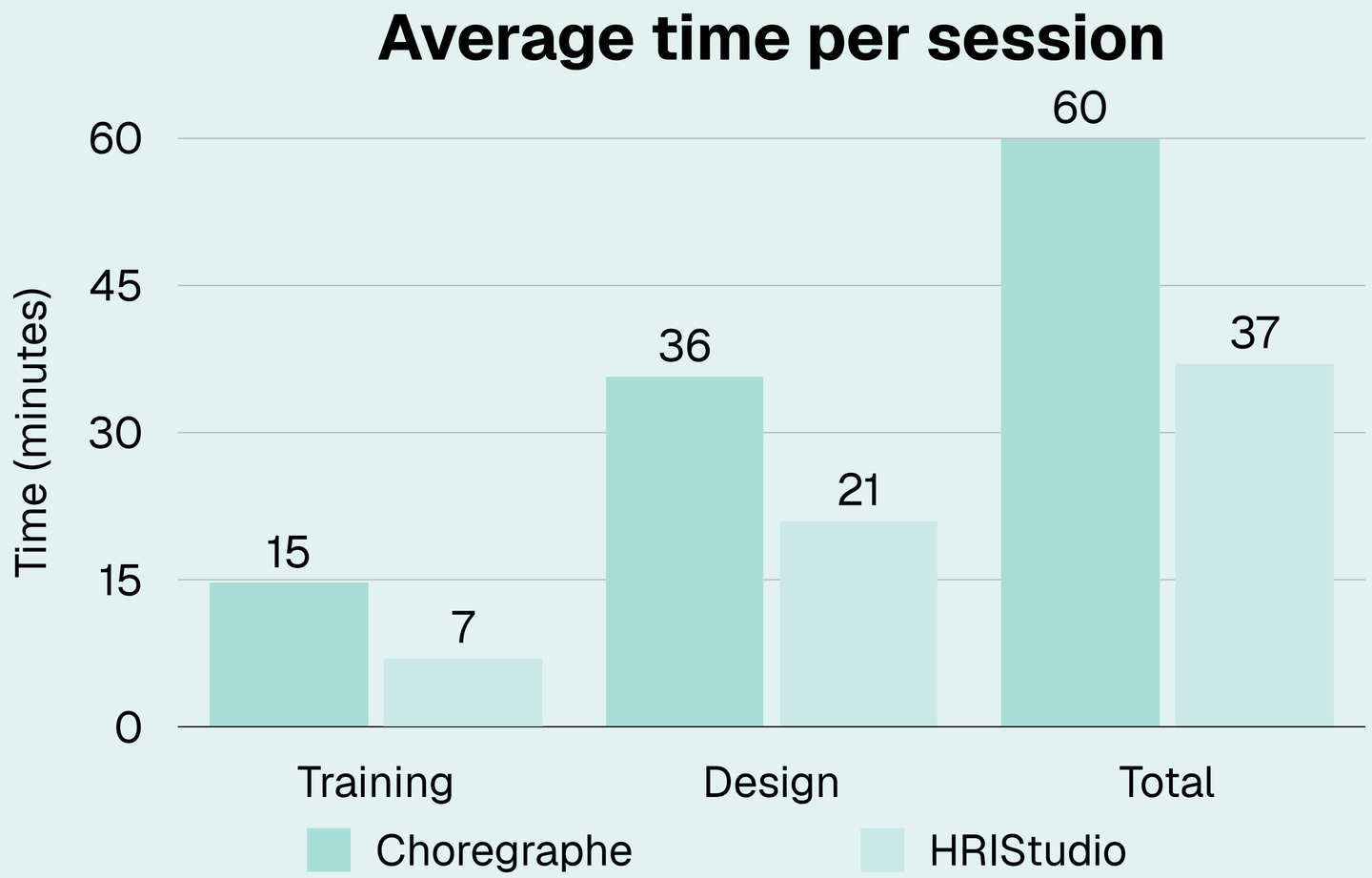
Reliable execution achieved.

System Usability Scale

77 points

Over the benchmark (>68)

Phase Timing



Tool Training Time

8 mins saved

Per session, on average.

Design Time

15 mins saved

Per session, on average.

Total Session Time

23 mins saved

Per session, on average

Selected Anecdotes from the Pilot

Accessibility in Action

- W-05 (HRIStudio): Zero programming experience. Perfect design in 18 minutes.
- W-04 (Choregraphe): Moderate programming experience. Incomplete design after 35 minutes.

The Silent Deviation

W-01 (Choregraphe): Accidentally changed a core script detail mid-trial.
The system allowed it and failed to log the error.

Conclusions

Conclusions

The *Accessibility Problem* - Answered

HRIStudio scored 76.7 on the SUS — above the 68-point benchmark — while Choregraphe scored 59.2. Non-programmers can design and run complete WoZ studies.

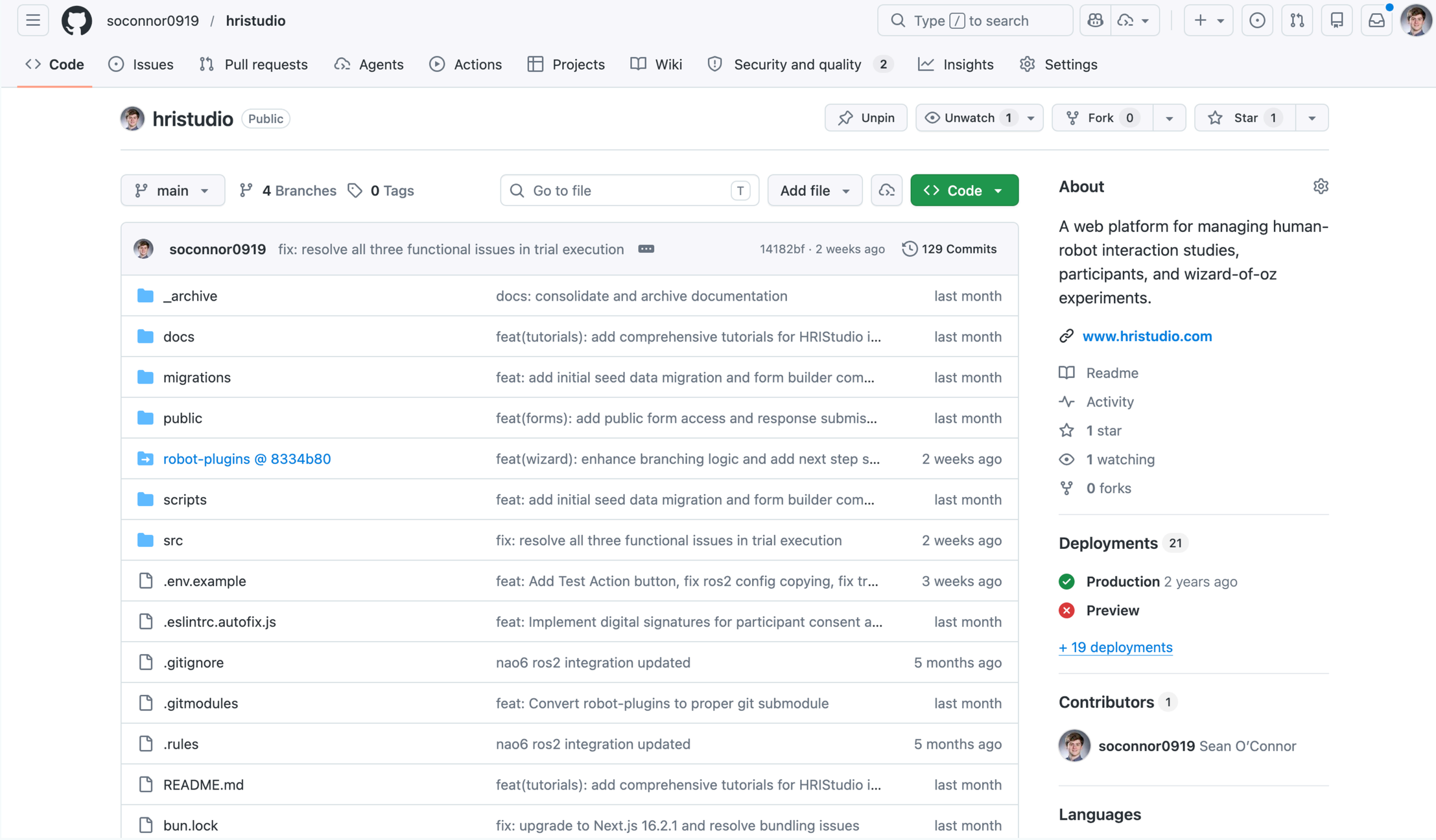
The *Reproducibility Problem* - Answered

HRIStudio achieved perfect design fidelity (DFS = 100) and near-perfect execution reliability (ERS = 96.7). Study designs can be faithfully replicated.

The contribution- not just an architecture, but an open platform for the field.

HRIStudio is open-source and robot-agnostic. Study designs are shareable, version-controllable files. The plugin architecture supports rapid addition of robot platforms. HRIStudio can be the basis upon which future studies are built- without having to engineer new tooling.

A Web-Based Wizard-of-Oz Platform for Collaborative and Reproducible Human-Robot Interaction Research



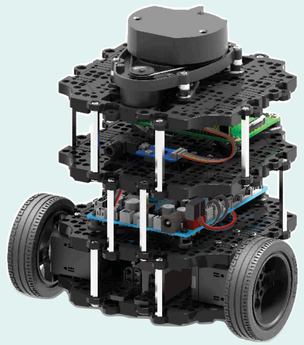
Future Work

Larger validation studies

The work done for this thesis is a pilot study- designed to prove feasibility of such a protocol and platform. With more time and funding, larger-scale studies can go more in depth.

Additional robot platform support

The architecture developed throughout this project has been designed to support multiple robotics platforms, yet in it’s current state, HRIStudio only supports the NAO. As demand changes, additional robot plugins can be implemented.



Thank you.

Any questions?